Against the Grain

Volume 23 | Issue 4 Article 11

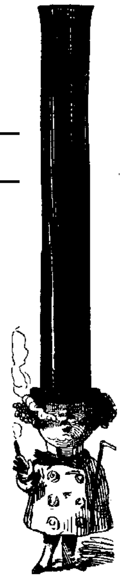
September 2011

Digital Archiving Challenges and Solutions: A Personal Perspective

Robert W. Boissy
Springer Science + Business Media

Digital Archiving Challenges and Solutions: A Personal Perspective

by **Robert W. Boissy** (Manager of Account Development & Strategic Alliances, Springer Science + Business Media)



Managing the Transition to Electronic Journals

In 2003, when I switched from technical services at a subscription agency to academic licensing for a publisher, the main conversation taking place with libraries was how and when to manage the switch to e-journals. One of the barriers for some libraries was uncertainty about the long-term reliability of e-journals. What steps were publishers taking to safeguard the legacy of scholarly publications? For some publishers, electronic publishing was still around the corner, and for others it was new enough that access was given away as an incentive to maintain print journal subscriptions. Many publishers had special deals to allow libraries to pay marginal additional amounts to maintain a subscription in both print and electronic formats.

Uppermost in the minds of librarians was the question of whether e-journals were a sound investment. Could the library own and possess the content in the same way as a print journal? A term was borrowed from divinity schools — "perpetual," as in perpetual access. Could libraries trust publishers, especially commercial publishers, to archive content safely and make it available to subscribers forever? This had always been a core responsibility of the library community and now it appeared to be shifting to the publishing community. Early library adopters of e-journals began to insist on licensing language relating to matters of ownership, archival rights, perpetual access, backup rights for local media storage, etc. License language provided incrementally more insurance for the e-journal purchasers, but how does legal language help if the publishing house is dissolved or otherwise unable to meet the terms?

In response to library concerns about long-term reliability, publishers established digital archives on equipment owned and maintained by trusted partners such as national libraries. This was somewhat reassuring, but not completely so. After all, the local library considering purchases of e-journals had no say in the legal or technical arrangements between publishers and national libraries. How would access work in the event of mishaps such as publisher server failures? Would national libraries make content accessible to libraries from other nations? Libraries and consortia searched for options and alternatives. Publishers aggressively pushing the transition from print to e-journals also cast about for additional solutions. They were eager to remove obstacles and gain cost savings from the reduction of physical issue logistics.

An early and clever solution was offered by the folks at **Stanford** in their **Lots of Copies Keep Stuff Safe (LOCKSS)** software for local backup of purchased content. The technical workings of **LOCKSS** are described on its Website (*http://lockss.stanford.edu/lockss/Home*), and this solution has been discussed extensively in the professional literature. The genius of **LOCKSS** is its ability to run the free software on inexpensive hardware and then step in virtually invisibly if access to some or all subscribed content on a publisher server is interrupted. It has taken some time for publishers to fully embrace the **LOCKSS** idea and open their servers to subscribers running incremental **LOCKSS** backups. Also, libraries are not uniformly interested in running local backup software. But it is a solution allowing libraries to possess their purchased content as they did in the print era. There is a valid school of thought holding that, if you want to do a job right, do it yourself. As a publisher representative, I always pushed for the **LOCKSS** approach because it took the pressure off my company to keep coming up with incrementally safer solutions with intermediaries.

Another option is a shared print archive. Some publishers have worked with consortia or state organizations to ship print copies of their content to long-term storage facilities under regional control of a group of libraries. Of course, this has the slight disadvantage of keeping publishers who wish to transition to electronic output in the print publishing business. Print archiving is a response to those who strongly associate safe long-term access with paper but who also acknowledge the convenience and utility of e-journals.

Commercial intermediary archiving solutions like **Portico** by **Ithaka** (*http://www.ithaka.org/portico*) have a strong following among both libraries and publishers. Libraries gain the advantage of having a paid, legal arrangement with a party other than the publisher of content to ensure long-term access to the content. Many libraries think of this as an insurance policy. Trigger events for making content available on **Portico** are well-documented, and many publishers participate.

Who Do You Trust?

Of course, the safety/risk equation really has no end, as any proposed solution could go wrong in one way or another. Perhaps it makes more sense to discuss which sector of the library and information community is best situated to take responsibility for the task of long-term preservation and access to digital content. Another way to think of who is best able to handle the task is to consider which organization or sector of organizations is most trusted by the most stakeholders because of its administration, constitution, governance, and resources. There is a good case to be made for one organization.

Controlled Lots of Copies Keep Stuff Safe (CLOCKSS) is a marriage of the local digital archiving idea with the national or large, trusted library idea. **CLOCKSS** (*http://www.clockss.org/clockss/Home*) is an organization managed for the good of the library and publishing community by representatives from these communities. A group of large member libraries with more staff and technical resources essentially backs up publisher content using **LOCKSS** technology into a secure dark (inaccessible) archive, with the understanding that they will be responsible to make that content available to everyone for free under certain well-described trigger conditions, including:

- The publisher is no longer in business;
- The content is no longer offered online from any publisher; or
- A catastrophic technical failure or natural disaster occurs.

In all cases, if a legal rights holder is able to maintain appropriate availability of the content, the **CLOCKSS** trigger does not apply. The representatives of the **CLOCKSS** Board are responsible for making these decisions. It is a joint effort; the balance is right. **CLOCKSS** is especially good because it strengthens the digital safety net for all libraries (and the public at large), shares responsibility between publishers and libraries, and lifts the burden of extensive redundant archiving from the general population of libraries. More backups equals more security, but **CLOCKSS** lifts the burden of carrying out these backups from libraries who do not have the staff to perform them or the money to pay for commercial backups. And, of course, it makes it easier for e-publishers to do what they want to do, which is eliminate as much print publishing as possible.

Publisher Software

A variation on the digital archiving problem is the publisher software problem. Publisher representatives hear two somewhat conflicting messages from libraries when it comes to long-term access to content. The first message is that publishers should not invest very much in the functionality of their own content platform because users do not generally start searches or even navigate much on publisher sites. Better to keep development and content costs down, the libraries say. The **International Consortium of Library Consortia (ICOLC)** has issued such a statement (*http://www.library.yale.edu/consortia/icolc-econcrisis-0109.htm*). It is undeniable that users often do not start their content searches on publisher Websites, but instead discover content from a handful of search engines, indexing and discovery services, referrals from teachers and colleagues, library catalogs, classroom software, other websites, etc. Of course, because usage drives renewal decisions and authors want their work to be widely available and read, it is in the publishers' interests to make their sites attractive and easy to use.

continued on page 38

Digital Archiving Challenges ...
from page 36

The second message that publishers hear from libraries is that local and intermediary services are not a perfect backup because they do not incorporate the publisher host system software; in other words, the libraries are saying that the browse, search, index, and navigation elements of publisher sites are important after all. Why should a library backup publisher content on servers or other media when the content would be cumbersome to use without the associated indexing and platform software? What seems to be wanting is an insurance policy that at least preserves the publisher software in the case of a disaster. Such a policy would allow for an orderly transition for the content onto some other platform. **CLOCKSS** and **Portico** offer their own solutions to this problem; basically, they revert, if necessary, over to their own systems and rely on a combination of the same popular indexing, discovery, and linking services that cover the current publisher content.

Keeping Tabs on Publishers

With physical collections, library buildings served two purposes: making content available for local use in a pleasant setting, and preserving that content. With digital collections, both those purposes are radically altered. So the question libraries ask themselves is whether it is still their job to safeguard the content. A logical question to ask is whether the content will be adequately preserved if the library chooses to focus elsewhere. This is an important question, but in a kind of distant, theoretical sense. Generally speaking, it appears that there are enough efforts underway and enough responsible parties stepping up to tackle this problem, that the preservation question is answered. Digital copies of published content are being stored in many places, in many countries, and on many different servers. And the same content is being stored in print archives by responsible initiatives. The responsible library task is therefore not necessarily to locally archive content, but to keep track of archiving by publishers and insist on information from publishers about their archiving initiatives. For information on developments in this area, follow the work of the **Piloting an E-Journals Preservation Registry Service (PEPRS)** project at **University of Edinburgh** (*http://edina.ac.uk/projects/peprs_summary.html*). The most important question for each purchasing library is therefore not whether the record of human publishing and achievement will be preserved perpetually, but whether content will be available to the local library's users.

So what are the potential obstacles to local access to paid content? The most likely is the inability to continue subscribing to content and the consequent possibility of a loss of back access. To safeguard against this outcome, libraries must negotiate into licenses fair clauses assuring back access in the event that the library is unable to pay for some or all of their subscribed content. There is not complete consensus on what is fair in this regard, but one view is that back access to previously paid content should continue on the publisher server for as long as the library continues to purchase at least some content from the publisher. From this perspective, it is only after the complete severance of a financial relationship that a reasonable "server maintenance" fee should be instituted by the publisher for ongoing access. Since publishers want the most access on their own servers as possible, it is unwise to redirect clients to another provider until absolutely necessary.

Another possible problem can arise when content shifts from one publisher to another. The Transfer Code of Conduct (*http://www.uksg.org/transfer*) addresses this problem. It is a voluntary promise made publicly among signatory publishers to safeguard libraries against loss of access to paid content during the shift of content from one rights holder to another. In general, access to paid content should never be lost because of a shift of content among publishers, but it pays to know which publishers are actually signatories to the Transfer Code. Inconvenience may happen in the form of needing to adjust a pointer or linking service to a different platform, but that is different than complete loss of access under threat of paying twice for the same content. Note that this differs from paying for digital archives of "born print" content, where the associated fees are for the value-added processes of digitization and aggregation.

continued on page 40

against the grain people profile

Robert W. Boissy

Manager, Account Development and Strategic Alliances
Springer Science+Business Media
233 Spring Street, New York, NY 10013
Phone: (781) 244-7918 • <Robert.Boissy@Springer.Com>
www.springer.com

BORN AND LIVED: Was born and raised in South Salem, NY, now settled in Plainville, MA.

EARLY LIFE: **Middlebury College**, BA. **SUNY Albany**, MLS. **Syracuse University**, Certificate of Advanced Study. Post-Graduate Internship at **IBM T.J. Watson Research Center**.

PROFESSIONAL CAREER AND ACTIVITIES: Various training, support, and data exchange roles at a subscription agency for 15 years, and various licensing and marketing roles in STM publishing for 8 years. Former Chair **International Committee for Electronic Data Interchange for Serials (ICEDIS)**, as of June 5th 2011 Vice-President/President-elect of the **North American Serials Interest Group (NASIG)**.

FAMILY: Married to wife **Kathy** for 25 years, children **Laura** (20), **Libby** (17), and **James** (15).

IN MY SPARE TIME: Reading essays, landscaping, watching amateur and professional soccer.

FAVORITE BOOKS: Recently, *The Thousand Autumns of Jacob de Zoet*, and *Cloud Atlas*, both by **David Mitchell**.

PET PEEVES: Getting older and slower as the world moves faster and faster.

PHILOSOPHY: Summed up nicely by **Bill McKibben** and his 350.org movement. Stewardship of the planet.

MOST MEMORABLE CAREER ACHIEVEMENT: Being elected VP/President-elect of **NASIG** by my peers in the serials community.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Learn how to travel gracefully and selectively after 25 years of running from place to place.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: Market forces have been driving many changes to the way scholarship is disseminated, and they will continue to redistribute costs and value as appropriate. Reading on tablets turns out to be fun, but all hardware and all electronic formats are fleeting. The need to preserve resources of all kinds in permanent, trusted storage mechanisms will begin to dominate the landscape in five years. Data mining will be honed to a point where hypothesis generation will be possible.

Digital Archiving Challenges ...
from page 38

Catastrophic failure of a publisher server is a far less likely obstacle to access. The raw content seems well covered by the network of archiving initiatives already described. But the temporary provision of publisher site software seems like a weak spot, especially for backup of content on local media. For this reason, it is fair to ask if a publisher has its own mirror site(s) or backup site(s) at a location far from its main servers. As the mass of data accumulates on ever larger publisher servers, this is an important topic of discussion for the experts in archiving like **CLOCKSS** libraries, national libraries, and major organizations serving library and publisher communities.

Archiving discussions always seem to assume a "big publisher" dimension, but they are perhaps even more important to small publishers and open access publishers. To the extent that a publisher is actively involved in electronic publishing, they need to have an archiving plan, and they need to be willing to share the plan's details with clients. Platform providers may take on a larger role for these kinds of tasks for smaller publishers, which is fine as long as the platform provider is responsive to the library community.

Finally, we remind ourselves that the lessons learned about digital archiving from the transition to e-journals also apply to the ongoing transition to electronic books. As eBooks continue to gain popularity, the time is right to settle these archiving matters. The next difficult steps will be in the area of improved archiving solutions for content that is dynamic, integrated, interlinked, and constantly updated.

against the grain people profile

Phoebe Ayers

Reference, Collections and Instruction Librarian
Physical Sciences & Engineering Library, UC Davis
Phys. Sci. & Eng. Library, 1 Shields Avenue, UC Davis, Davis CA 95616
Phone: (530) 752-9948 • Fax: (530) 752-4719
<psayers@ucdavis.edu> • *http://phoebeayers.info*

BORN AND LIVED: I was born and raised in Arkansas and moved to the West Coast in my late teens; I have lived in California and Washington State.

EARLY LIFE: I received a BA in English from the **University of Washington**, and also went to **UW** for library school; I received my MLIS in 2005. I have worked in a variety of libraries and a fisheries research lab, but discovered my love for engineering information as a graduate reference assistant at the **UW Engineering Library**.

PROFESSIONAL CAREER AND ACTIVITIES: At **UC Davis** I am the liaison librarian for computer science and electrical engineering. I work with faculty and students, work at the reference desk, and collect for these areas. I am also involved in open access outreach and data management activities. This year I am the president of our local SLA chapter (the Sierra Nevada chapter), which covers the Sacramento, CA and northern Nevada areas. I am also involved in **ACRL**, as well as serving as the Web manager for the **UC**-wide librarians association. I enjoy doing program planning and have served in this role for a variety of groups. Finally, I am a member of the **Wikimedia Foundation** Board of Trustees.

FAMILY: I have a wonderful extended family that is scattered all over the U.S., as well as a great network of friends that I have made over the years, so you will often find me visiting one corner of the country or another on the weekends. At home it's just me and some severely neglected houseplants.

IN MY SPARE TIME: I love to travel, which is fortunate because I travel a lot for **Wikimedia** and library conferences, as well as personal trips. **Wikimedia** activities take up a good deal of my spare time; with the remainder I read, cook, make crafts, ride my bicycle, watch bad science fiction TV, and spend way too much time on the Internet.

Something to Think About — Doubling Up?!

Column Editor: **Mary E. (Tinker) Massey** (Retired, Serials Librarian, Embry-Riddle Aeronautical University, Jack R. Hunt Library) <eileen4tinker@yahoo.com>

Having a little time to think about the present and future of librarianship, I looked at new job vacancies to see what the demands of the field are. We are still facing drastic cuts in operating funds, cutbacks on numbers of positions, and demands to reformat our functioning organizational structure. We have to sit down and figure out what services are more necessary than others and reallocate our workforce to take care of those changes. Granted, new technology has caused us to re-evaluate how we operate for our patrons. How can we begin to make some sense of it? Perhaps we should consider encouraging our staff to further their education by acquiring other certifications and/or degrees to add to their abilities. We are used to having a few librarians increase the number of their credentials in specific areas, such as Music, Engineering, or some other appropriate field. We are now assessing jobs and coupling some of the tasks in order to make things work. The problem becomes the fact that people are being asked to do jobs that they are not exactly trained to do. You may have training to catalog monographs, but you may need more training to catalog serials or media or documents. Libraries will have to put some of their monies into retraining and furthering education for their present staff. But we must also re-evaluate those needs properly and get the best bang for our bucks. Sending staff to conferences, training workshops, and virtual sessions to update their credentials has become essential — not a luxury. In our small institution alone, we have found a 95% increase in those staff who are now engaged in advanced training or retraining activities. I have been impressed by this increase and hopeful that these staff members will be the ones to retrain others on our staff.

Increased knowledge will have to be obtained in preservation techniques and digital preservation to maintain the viability of our collections and rare materials. That process has begun and soon there will be grants formulated to accomplish many of the dreams we have had. The library has invited me back in the future to see their results on one of my basic passions I fought to establish over the six years I worked there. I am excited to still be a part of this.

I guess the doubling up I speak about reminds us to keep improving our knowledge in many areas, but it also insinuates that we should be backing up our positions with others who also understand the needs and tasks and can operate on them when the primary person is not there. I have seen too many cases, in both small and large libraries, where only one person knows the tasks and has been out on extended family leave, personal illnesses, or accidents. We can barely function on the reduced staff now, so cross-training is essential.

I think doubling up is indeed something to think about? What say you? Get involved in your library to help that change occur!